June 2024

Exploring Topic-Related User Experiences Through Social Graph

Shelly Gupta
Temple University, shelly.gupta@temple.edu

Ameen Abdel Hai
Temple University, aabdelhai@temple.edu

Abdulrahman Alharbi
Temple University, abdulrahman.alharbi0001@temple.edu

Hussain Otudi
Temple University, hussain.otudi@temple.edu

Zoran Obradovic
Temple University, zoran.obradovic@temple.edu

EXPLORING TOPIC-RELATED USER EXPERIENCES THROUGH SOCIAL GRAPH

Completed Research Paper

Shelly Gupta, Temple University, Philadelphia, USA, shelly.gupta@temple.edu
Ameen Abdel Hai, Temple University, Philadelphia, USA, aabdelhai@temple.edu
Abdulrahman Alharbi, Temple University, Philadelphia, USA, adalharbi@temple.edu
Hussain Otudi, Temple University, Philadelphia, USA, hussain.otudi@temple.edu
Zoran Obradovic, Temple University, Philadelphia, USA, zoran.obradovic@temple.edu

Abstract

Understanding user experience presents challenges for researchers since gathering user feedback can be costly and time-consuming. This study introduces a novel semi-automated framework for analyzing user experiences with a topic on Reddit. The framework identifies relevant subreddits based on a seed list and then explores users' histories to find topically related communities. A heterogeneous graph is constructed to represent interactions between users and subreddits, with users' activities captured as BERT-generated textual node vectors. The proposed framework is evaluated using a use case of Pay-per-Click technology. Furthermore, the versatility of the proposed data collection method is demonstrated in a news analytics application related to the Russia-Ukraine conflict using the New York Times articles and comments. Results show that the inclusion of neighborhood subreddits significantly broadens the scope of user topic analysis. Furthermore, the proposed graph structure-based framework outperformed four alternatives considered providing evidence that it can effectively predict missing user-subreddit and user-user interactions.

Keywords: Social media, Graph modeling, Pay-per-Click, Link prediction.

1 Introduction

Understanding user experience is a focus of numerous research endeavors, as it offers insights into the intricate dynamics between users and topic. This comprehension is important because it provides a holistic understanding of users' feelings, perceptions, and interactions with a particular topic. Investigating user experience is vital in assessing the effectiveness of a technology, as it allows for the identification of user preferences, and areas for improvement (Law and Van Schaik, 2010). Traditional approaches to investigate user experiences have often entailed labor-intensive and cost-prohibitive methods such as participant recruitment and survey administration (L. Zhu and Lv, 2023). While these conventional approaches have yielded significant findings, their inherent limitations include time consumption and costs.

An efficient and cost-friendly method of capturing users' interactions is to gather openly available data from social media platforms. Platforms like Reddit see increased engagement as specific topics gain popularity, with users discussing and sharing experiences related to the topic (Prabowo et al., 2021). Reddit, a widely used online discussion platform, consists of numerous subreddits covering various topics (Medvedev, Lambiotte, and Delvenne, 2019). In this regard, Reddit has emerged as a central hub for users to exchange knowledge and insights regarding a topic, making it an appropriate platform for studying the interactions and dynamics within an array of particular topic-related communities.

When utilizing social media platforms such as Reddit, data collection is a struggle as it can be a time-consuming task to locate relevant posts about a technology. Furthermore, the data contains missing information and multiple informal contractions (Proferes et al., 2021). Additionally, the enormous volumes of data present challenges when attempting to model user interactions.

This study addresses these gaps by introducing an end-to-end semi-automated framework that tackles the above-mentioned shortcoming and utilizes Reddit to generate a user-subreddit interaction graph for a particular topic. In the proposed approach, the user begins by inputting a list of seed subreddits associated with a topic. Subsequently, a neighborhood is examined to identify the related one-hop subreddit neighbors and aggregate the posts and comments from all the identified subreddits. Extensive prepossessing and data mining techniques are utilized to aid the model in learning a pattern. Finally, a comprehensive graph is generated that consists of subreddits and users as nodes, with user interactions represented as edges thus providing means for assessing user interactions related to a topic. Specifically, this work aims to answer the following research questions:

- **(RQ 1)** Is there a topical value added to the graph structure by the inclusion of neighborhood information or "one-hop neighbor" subreddits to the graph? The assessment involves examining the overlap in the topics discussed on original subreddits and neighborhood subreddits, while also identifying any new topics that emerge from the neighborhood discussions.
- **(RQ 2)** To what extent is the information retained within the proposed graph that includes one hop neighbor information? Can this graph predict missing links better than the baseline graphing methods?
- **(RQ 3)** How much predictive value does contextual information contribute to the performance of the graph in the link prediction task?

Topic modeling, word cloud analysis, and exploration of topological features of the graph are performed to answer RQ 1. Link prediction is a standard way of testing how well the information is stored in it and the benefits of considering neighborhood information when predicting links are evaluated to address RQ 2 and RQ 3. BERT-based contextual embeddings are crucial to the framework, and RQ 3 aims to evaluate their advantages. The contributions of this study are summarized as follows:

- A novel graph-based framework is proposed to model relationships between users for a particular topic.
- An end-to-end semi-automated framework is created that takes a set of seed subreddits as input and generates a subreddits-users interaction graph.
- The work presents a semi-automated data collection framework that explores the neighborhood for the seed subreddits and finds topically relevant neighborhood subreddits or "one-hop neighbors".
- Extensive experimentation is conducted to assess the hypothesis that the proposed neighborhood aggregation methods add contextual value to user experience analysis.
- Link prediction is conducted to assess and evaluate the network's ability to successfully capture the interactions in the graph.

This study demonstrates the efficacy of the proposed framework using a use case of Pay-per-Click (PPC) technology. PPC represents a cutting-edge approach to tailor advertisements for specific audiences, where advertisers pay based on the number of clicks on their brands links (Kapoor, Dwivedi, and Piercy, 2016). While machine learning is extensively employed in digital marketing for tasks like ad customization (Krishen et al., 2021), there remains a noticeable lack of research dedicated to harnessing PPC user experiences for informed decision-making. Consequently, PPC serves as an exemplary case study for evaluating the suitability of the proposed data collection and graph generation methodology. Moreover, this study delves into the generalization capability of the proposed framework by examining user comments on articles from The New York Times about the Russia-Ukraine conflict. News articles have long been a subject of study for analyzing reader comments and attitudes (Alshehri et al., 2021). They offer a valuable resource for evaluating the effectiveness of the proposed data collection method. By scrutinizing

comments on news articles, the study aims to assess how well the approach adapts to diverse sources and topics.

2 Related works

Methods of collecting social media data: The idea of using Reddit as a data source is bolstered by the ease of use of the official Pushift API (Baumgartner et al., 2020) which is available for public use. Multiple methods have been used to collect data from Reddit in the past. Numerous studies have used pre-made datasets (Benslimane et al., 2023) but for niche problems, such datasets are hard to come by. Many studies rely on collecting data by performing a keyword search and gathering the results (Jiang et al., 2020) but when dealing with a broad topic like "digital marketing" or "pay-per-click", keyword generation is a hard and expensive task that involves a great amount of human intervention. Some studies rely on collecting all the data from manually curated subreddits' list (Boettcher, 2021) but studies have shown that highly correlated subreddits exist and users actively participate in these "spinoff" or "complementary" communities as well (Hessel, Tan, and Lee, 2016). Hence, to fully understand a user's interaction with an entity, a wholesome review is needed to identify these communities which may not be obvious in manual curation. Keeping this in mind, the proposed framework collects not only the data from seed subreddits but also explores users' activity to find the topically similar communities and include them in the dataset.

Different ways of creating social user graph: Different studies have visualized Reddit discussions in heterogeneous ways. Some studies model Reddit as a" post-to- post network" where posts share a link if they were posted by the same user (Hurtado, Ray, and Marculescu, 2019). This method albeit its advantages, fails to include comments in the graph. Comments are a rich source of information but are often overlooked as they exponentially increase the raw data and make processes computationally expensive. Some studies that take comments into account rely on a separate graph for each post which contains comments as nodes (Benslimane et al., 2023). This method is flawed in the sense that it is not scalable. Many subreddits contain millions of posts and comments and crafting an individual graph for each post is not sustainable. One study focused on modeling 2 types of graphs: "User-subreddit-User" network where a link between users exists if they share greater than 7 subreddits and "Subreddit-user-Subreddit" network where links are established if the communities share at least one common user (Janchevski and Gievska, 2019). Rather than having two different graphs, the proposed approach combines users and subreddits nodes in one heterogeneous graph which is more inclusive due to lack of arbitrary thresholds. One study models activity overall by creating a user's and subreddits nodes-based graph (Aleksic et al., 2022) but this study does not take comments and textual information into account.

Need for social network analysis for PPC: It has been well established that Word of Mouth (WoM) greatly impacts the purchasing decisions of consumers and affects how a user perceives their relationship with a brand (Yang, Cheng, and Tong, 2015). Numerous studies have explored social media and determined that it not only increases the pervasiveness of WoM but also its impact on the public (Alalwan et al., 2017). Today, more people are discussing their positive or negative experiences with a brand or technology on social media due to the presence of large communities on these platforms and the sense of empowerment felt by the user (Hudson et al., 2016). Therefore, a study of social sensors is the next step when trying to analyze and understand users' experience with digital marketing and Pay-per-Click technology.

Myriad of researchers have focused on studying electronic WoM on social media platforms like Twitter and Facebook but Reddit as a data source seems relatively less explored. Twitter and Facebook data is shorter and considered "on the fly" i.e., users are posting about events as they unfold. In contrast, Reddit provides longer and contextually richer texts. The users on Reddit are encouraged to post with a "cool head" (Bonifazi et al., 2023). This aids in achieving a qualitatively enhanced analysis. Reddit has a large user base and data is arranged as part of topic-specific communities which makes relevant data extraction simpler (Amaya et al., 2021). Therefore, in this study, Reddit is explored as the data source.

Social network-based studies for digital marketing: Various studies have explored natural language

processing research interests in the field of digital marketing like customer targeting and brand positioning (Shankar and Parsana, 2022). One article analyzes brand communities on Twitter by creating a heterogeneous graph out of users, their posts, hashtags, and users mentioned in the post (Brambilla and Gasparini, 2019). But right now, Reddit is yet to be fully explored to determine the relationship between users and digital marketing platforms. One study aggregated posts and comments per user to identify user types and segmentation in online brand communities (Ge, Zhao, and Zhang, 2022). The graphical representation is not considered. Another work creates a user-user graph for" r/intel" and discusses a type of posts present in the community (Oliveira and Marques dos Santos, 2022). However, textual information is not added to the graph and is studied separately. Moreover, the discussion of the graph does not go beyond basic network properties like density.

The aforementioned studies propose methods that cannot capture a broad view of a topic on Reddit. These methods are either limited by their choice of dataset collection method or they limit the focus of their research to Reddit posts only. These graph generation methods are unscalable or lacking comment information. The proposed framework is designed to overcome these gaps. It requires minimum human intervention and therefore, mitigates costs, improves efficiency, and improves understanding of user experience.

3 Methodology

The proposed semi-automatic framework has been visually depicted in Figure 1. It comprises of three steps: 1) Data collection; 2) Data preprocessing; and 3) Graph modeling. The steps are illustrated in detail in the following subsections.

3.1 Data collection

This study introduces an innovative approach to collecting Reddit data. Initially, users are asked to select "seed" subreddits related to a specific topic and specify a timeframe of interest. Two distinct Reddit APIs PRAW and the Pushshift API are employed for the retrieval of comments and posts, respectively. Following the initial data retrieval, the incorporation of the neighborhood is proposed to further expand the dataset and gain a more comprehensive understanding of the technology by broadening the contextual perspective. This neighborhood expansion enables the proposed framework to identify all subreddits where these users are active, commonly referred to as 'one-hop neighbors.' The framework performs neighborhood aggregation by collecting data from all users in the original seed subreddits and aggregates each user's posting history to compile a list of all neighborhood subreddit communities.

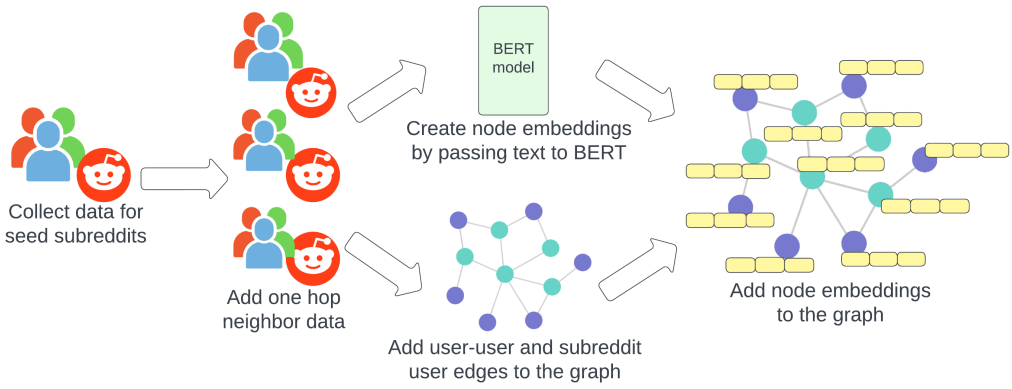

Figure 1. Proposed graph modeling framework. Purple nodes are subreddit nodes and cyan nodes denote users. Users and subreddits share links but subreddits are not connected.

To streamline the focus of the analysis, a threshold is applied for selecting subreddits with potential

relevance. Specifically, attention is narrowed down to subreddits where at least a given percentage of the users are actively participating, thus considering them as potentially topically relevant communities or "one-hop neighbors." Note that this threshold value is a hyper-parameter and should be adjusted for each use case. Multiple thresholds were experimented with during the PPC case study, and it was found that the 1% threshold yielded the best results for that particular use case. Similarly, for the New York Times application, the 0.75% threshold generated the most accurate results. After the threshold selection, user intervention is introduced to manually select the most relevant one-hop neighbors for further analysis. Subsequently, the proposed framework proceeds with the comprehensive collection of posts and comments from these selected top one-hop neighbors, effectively diversifying the dataset. It is important to emphasize that, due to the stringent moderation practices on Reddit, all data retrieved from these communities is considered as relevant and valuable for the research.

3.2 Data preprocessing

Social sensor data, in general, requires thorough cleaning for several reasons. Firstly, textual data often contains noise and additional information, such as emojis, hashtags, and special characters, which can impact the accuracy of the analysis. Secondly, different social media platforms have their own formats and jargon. Cleaning and standardizing the data help overcome these challenges. The following preprocessing steps were implemented as part of the proposed framework:

- During data analysis, it was discovered that a considerable number of posts were missing essential information such as post titles and author IDs. Such posts were excluded since these variables are critical for graph modeling. A similar approach was applied to the comments data as well to ensure consistency in the method.
- Many posts were missing 'self-text' information, and it was replaced by an empty string. 'Self-text' information denotes text that author might have added to the post apart from the title.
- All non-English posts and words were removed because such words might be interpreted as misspelled or unknown words, resulting in incorrect or inaccurate results for natural language processing (NLP) tasks.
- Social media data often includes contractions such as 'g8' for 'great,' which are not compatible with BERT. To overcome this issue, the framework utilized a Python library known as 'contractions' to expand known contractions and transform the Reddit posts into a more formal language.

3.3 Graph modeling

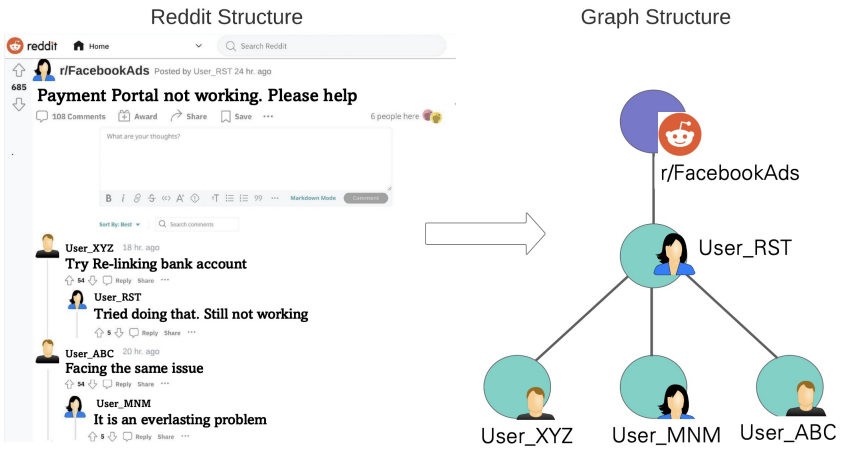

Figure 2. Converting subreddit posts into a graph. The comment tree structure is flattened out and the self-loop has been removed.

The framework constructs a graph representing the interactions between users and subreddits related to the topic by including nodes of two different types. First type of nodes represents all the users who created a post or a comment in the subreddits. Second, the subreddits themselves are also included as nodes. These two types of nodes help understand the pattern of information flow between users and subreddits. The hypothesis is that this heterogeneous graph-based approach provides a more comprehensive understanding of the structure and evolution of the topic's community.

The graph construction process involves adding different types of edges between nodes. Specifically, two types of edges were incorporated into the network: user-subreddit edges and user-user edges. User-subreddit edges were used to link each user to the subreddit where they created at least one post. On the other hand, user-user edges were added to connect users who interacted with each other through comments on their respective posts (refer to Figure 2). It is important to note that Reddit organizes comments in a tree structure, but for the proposed graphical representation, the comment tree was flattened out, i.e., the hierarchical information was excluded. This decision is made because the focus is on capturing user-to-user connections rather than the hierarchical structure of comments. The aim is to depict how users engage with specific posts and indicate the likelihood of a user commenting on another user's post. Additionally, self-loops were not considered as they do not provide meaningful information for the analysis and link prediction task.

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained deep learning model with a transformer-based architecture. It can generate text embeddings that are contextualized, meaning that the embedding of a word can change based on its context (Devlin et al., 2018). In this framework, the pretrained BERT model is implemented to create textual embeddings. The creation of user node attributes involves concatenating all the user's posts and comments, followed by generating BERT embeddings. Subreddit node embeddings, on the other hand, are generated by combining all the posts within a given subreddit and feeding them into the text vectorization model. Maximum BERT token size of 512 was used to create textual embeddings.

4 Case Study

Marketing is an ever-evolving and dynamic business activity that has transformed digital marketing alongside the rise of electronics. By harnessing technology, digital marketing aims to enhance marketing efforts and improve customer acquisition by determining designed demographics and usage characteristics (Desai and Vidyapeeth, 2019). One notable revolution in digital advertising has been the introduction of Pay-per-Click (PPC) technology. This innovative approach enables advertisers to create personalized ads intended for their targeted audience. These ads are then displayed on various platforms, with the advertiser being charged only when a user clicks on the ad (Kapoor, Dwivedi, and Piercy, 2016). The popularity of this model has surged due to its cost-effectiveness and its ability to generate high click ratios (Agarwal, 2021). The advancement of this domain can be attributed to the integration of big data and academic research in intelligent systems (Gkikas and Theodoridis, 2019). Numerous studies have concentrated on developing personalized ads by examining user behavior (Flaherty, Domegan, and Anand, 2021). Furthermore, significant attention has been given to the aesthetics of ad design (Y. Zhu et al., 2023) and the prediction of click fraud (Alzahrani and Aljabri, 2022). However, despite the widespread utilization of machine learning in the field of digital marketing, there is still a lack of research focused on leveraging PPC user experience for advanced decision-making. Hence, it serves as an appropriate case study to evaluate the effectiveness of the proposed methodology.

For this study, nine seed subreddits related to PPC ('r/googleads', 'r/adwords', 'r/FacebookAds', 'r/adops', 'r/redditads', 'r/bingads', 'r/Adsense', 'r/admob' and 'r/AdmobAds') were carefully chosen. In total, 61,922 posts and 180,289 comments were collected from these seed subreddits. Once the posts were obtained, their unique post IDs were used to retrieve all associated comments for each post. In the seed subreddits, there were a total of 26,269 unique authors, and the user-based neighborhood analysis revealed

a substantial 61,304 subreddit one-hop neighbors. To streamline the analysis, 1% user threshold was applied, reducing the number of one-hop neighbors to 220 subreddits. At this stage, user intervention was introduced to manually select the top 27 one-hop neighbors most relevant to the analysis. The original seed subreddits and the selected one-hop neighbors are listed in Table 1. Experimentation with multiples thresholds was conducted during use case study and it was found that 1% acts as the best. 0.5% threshold yielded 677 subreddits which would require some considerable manual effort during selection. 1.5% threshold resulted in 120 subreddits but the neighborhood was missing important relevant subreddits namely 'r/AffiliateMarketing', 'r/woocommerce', 'r/content_marketing' and 'r/InstagramMarketing'. Thus, 1% threshold served as the best tradeoff between finding potentially relevant one hop neighbor subreddits and the reduction of manual effort needed to select neighborhood subreddits. After incorporating information from the 27 one-hop neighbors, the expanded dataset encompasses 4,271,890 posts and 8,719,303 comments.

Seed subreddits	One Hop Neighbor Subreddits
'r/googleads,' 'r/adwords,' 'r/AdmobAds', 'r/adops,' 'r/redditads,' 'r/bingads,' 'r/Adsense,' 'r/admob,' 'r/FacebookAds'	'r/InstagramMarketing','r/buildapc' , 'r/marketing', 'r/webdev', 'r/shopify', 'r/ecommerce', 'r/RedditforBusiness', 'r/DigitalMarketing', 'r/digital_marketing', 'r/startups','r/dropship', 'r/AffiliateMarketing', 'r/AskMarketing', 'r/bigseo', 'r/SocialMediaMarketing','r/Emailmarketing', 'r/webhosting', 'r/graphic_design', 'r/FacebookAdvertising', 'r/woocommerce', 'r/content_marketing','r/PPC', 'r/SEO'

Table 1. Original seed subreddits and the selected one-hop neighbors after neighborhood aggregation for PPC.

5 Experimental setup

5.1 Inclusion of neighborhood information

A pivotal aspect of the framework revolves around the incorporation of one-hop neighbors' information into the dataset. An investigation into the impact of neighborhood analysis on the study was conducted. To comprehensively understand the effects, a series of three experiments was designed:

- **Topic Modeling:** In this experiment, BERTopic (Grootendorst, 2022) was employed to generate topics for each of the seed subreddits and their corresponding one-hop neighbors. BERTopic is a topic modeling approach that augmented the transformer-based models by the contextual Term Frequency-Inverse Document Frequency (c-TF-IDF). It yields hundreds of clusters (Topics), therefore enhancing the interpretability of resulting topics. Furthermore, it retains the prominence of critical terms within the descriptions of these topics. In this experiment, the top ten generated topics were used to perform deep analysis and were manually interpreted. The primary objective was to check for the overall overlap of topics between the seed subreddits and one-hop neighbors. This analysis allowed the reader to gauge the extent to which the inclusion of one-hop neighbors contributes additional information. Furthermore, the evaluation also delved into the emergence of new discussion topics, evaluating their relevance to the realm of PPC technology.
- **Word Cloud Analysis:** Word clouds are visual representations of the frequency of words in text. In this experiment, word clouds were crafted for the seed subreddits and one-hop neighbors. The intention was to gain insights into the nature of discussions occurring within these distinct communities. The examination of word clouds aimed to determine if these discussions were aligned with user experiences related to PPC technology.
- **Graph Properties:** In this experiment, the structural configuration of the proposed graph, which included data from the seed subreddits, and their one-hop neighbors was compared with a graph that solely contained data from the 9 seed subreddits. The objective was to analyze the differences and commonalities in the topological structure of the two graphs, providing valuable insights

into the network's evolution with the integration of one-hop neighbors' data. Specifically, the evaluation focuses on the number of nodes and edges, density, number of components, average degree, and average clustering coefficient.

Through these experiments, the aim was to illuminate the multifaceted impact of including one-hop neighbors' information, ultimately enhancing the depth and scope of the analysis within the PPC technology domain.

5.2 Link prediction

Link prediction task for graphs is a standard way of testing how well the information is stored in it. In an ideal situation, analysis should be able to reveal any missing links in the graph and/or predict future user interactions within the community. This link prediction problem was modeled as a binary classification task. The data was randomly divided into three parts: train (85%), validation (5%), test (10%), and equal number of negative links were added. Since the proposed network is a heterogeneous graph that contains two types of nodes and edges, performance was judged based on overall accuracy and accuracy for the following two objectives:

- **Subreddit-user link prediction:** the links capture the ability of the model to predict the likelihood of a user posting in each subreddit.
- **User-user link prediction:** These links capture the interaction between users. The model should be able to predict which users are more likely to communicate with each other than the rest.

The study utilizes a Graph Neural Network (Gori, Monfardini, and Scarselli, 2005) for aggregating graphs and conducting link prediction. This model comprises an encoder and decoder. The encoder transforms the input graph into a low-dimensional embedding, while the decoder generates output based on this embedding. A two-layer graph convolutional network is employed to generate node embeddings, encoding both nodes linked, concatenating their embeddings, and decoding for prediction.

BERT, a context-aware model for word embeddings, is powerful yet resource-intensive. It is prudent to know how much better results are achieved by adding the overhead of using BERT. Therefore, BERT's performance is compared with that of widely used FastText and Dov2vec embedding models. FastText is a lightweight word embedding model that has been pre-trained by Facebook and is openly available for public use (Bojanowski et al., 2017). Since FastText only generates word embeddings, document vectors are created by collecting embeddings of all the words present in the text and calculating their mean. On the other hand, Doc2Vec takes in documents of any size and calculates a numeric representation of the same has a sustainable impact on a model's performance (Le and Mikolov, 2014). The model is trained on Reddit posts and comments, serving as a baseline. Furthermore, BERT's performance is compared with other large language models like BERTweet and RoBERTa. BERTweet is an extension of BERT that has been fine-tuned on a large corpus of Twitter text (D. Q. Nguyen, Vu, and A. T. Nguyen, 2020). RoBERTa is a refinement of the BERT model that addresses certain limitations of BERT by removing next sentence prediction objective and is pretrained on a larger corpus (Liu et al., 2019). Token size of 512 was used for RoBERTa and 128 for BERTweet.

6 Results

6.1 Is there a value gained from conducting a neighborhood analysis of seed subreddits? (RQ 1)

Figure 3 provides a visual representation of the BERTopic-generated topics for both the seed subreddits and their corresponding one-hop neighbors, considered separately. The analysis reveals a substantial thematic connection to pay-per-click (PPC) technology in most topics. These topics shed light on various facets of PPC technology, encompassing areas such as click analysis, audience targeting, digital marketing,

influencer marketing, and marketing on diverse social media platforms like TikTok, Instagram, Facebook, Reddit, and others. The emergence of these insightful topics highlights the community's engagement with PPC technology and its various dimensions.

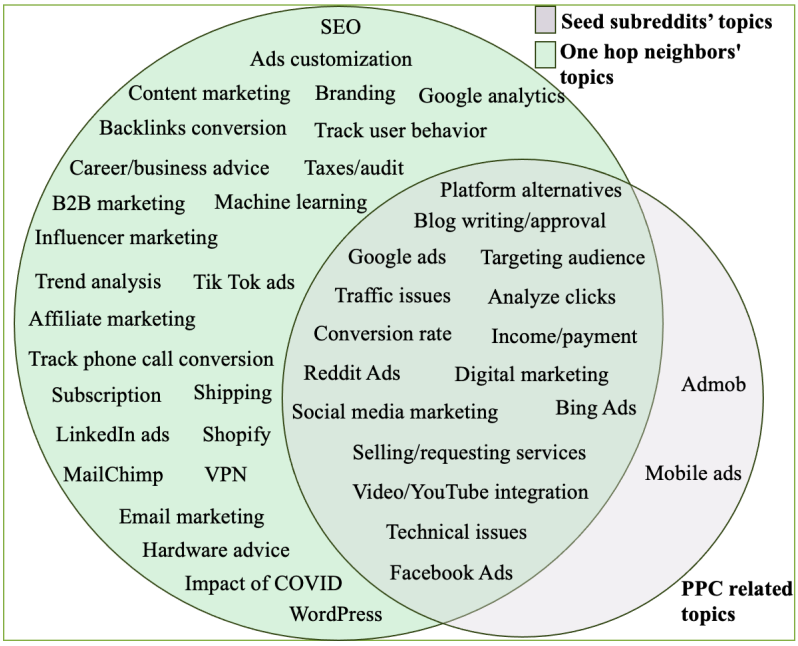

Figure 3. BERTopic topics generated for seed subreddits and one-hop neighbors. The topics have been manually interpreted.

Moreover, a noticeable alignment between the two sets of topics becomes apparent upon closer examination. Except for only two topics, the topics originating from the seed subreddits are also prevalent in their one-hop neighborhoods. Notably, topics such as "platform alternatives" in which users discuss alternative platforms to achieve similar goals, were found in both the seed subreddits ('r/adsense' and 'r/adops') and their one-hop neighbors ('r/dropship', 'r/emailmarketing' and 'r/webhosting'). This finding highlights the shared interest among users in exploring diverse platforms to meet their objectives, bridging the gap between the seed subreddits and their surrounding neighborhoods. Furthermore, users from both the seed subreddits and one-hop neighbors exhibited a distinct fascination with the integration of videos and YouTube into their content. They frequently engaged in discussions regarding technical challenges they encountered, demonstrating a shared concern for addressing such issues effectively. Additionally, blog writing and approval emerged as another common interest within this group, illustrating a mutual affinity for content creation and publication. In summary, the comprehensive analysis reinforces the notion that there is a clear overlap in the topics under discussion within the original subreddits and their neighborhood subreddits. This overlap signifies the coherence and shared interests among users across these communities.

Turning the attention to the topics that surfaced exclusively because of the inclusion of one-hop neighbors, new topics encompassing areas such as "ad customization", "branding", "Tik Tok ads" and "Shopify" emerge from one hop neighborhood which had remained conspicuously absent in the original discussions held within the seed subreddits. The inclusion of one-hop neighbors has, therefore, significantly broadened the conversational landscape by introducing fresh perspectives and dimensions to the analysis. It is particularly noteworthy that the seed subreddits, which were initially conceived with a primary focus on platforms like Facebook and Google, were inherently limited in their scope. However, the incorporation of one-hop neighbors led to inclusion of alternate platforms like LinkedIn, Instagram, WordPress, and others. This development showcases the dynamics of the online community and its adaptability in embracing a multitude of platforms.

Figure 4. Word clouds for seed subreddits (left) and one-hop neighbors (right).

Intriguingly, one of the novel topics that emerged pertains to the impact of COVID-19. This signifies a candid and real-time discussion among users regarding how the pandemic has influenced their businesses and the strategies they have employed to address the challenges posed by this global event. It becomes evident that users are sharing their real-life experiences and addressing the practical implications of such a significant event on Reddit. This illustrates that Reddit serves as a platform where users candidly discuss their day-to-day experiences, thereby reinforcing the assertion that the inclusion of neighborhood subreddits significantly broadens the scope of user analysis for a particular technology.

In Figure 4, the word clouds generated for both the seed subreddits, and their one-hop neighbors are presented, offering an intriguing glimpse into the dynamics of these communities and the nature of their discussions. A notable distinction in focus emerges between the two groups. The word cloud of the seed subreddits is distinctly concentrated on specific platforms such as Google and Facebook. Within this cloud, keywords like "help," "know," "looking," "issue," and "working" showcase the nature of discussions within these communities. These keywords show that users primarily use the seed subreddits as a platform to address the issues they encounter, seek assistance, and share their own experiences.

Conversely, when the attention is focused on the word cloud of the one-hop neighbors, a more generic and user experience-centric orientation is observed. While the cloud features fewer distinct keywords, it is rich in user experience descriptors such as ‘know’, ‘build’, ‘problem’, ‘want’, ‘work’, ‘looking’, and ‘wondering’. This reveals that the one-hop neighbor communities engage in discussions of a more general nature, with an emphasis on sharing and describing user experiences. It is worth noting that despite the differing emphases, a substantial overlap exists between the words that appear in both word clouds. This overlap signifies a common thread that runs through these communities, wherein the seed subreddits maintain a specific focus on the platforms, while the one-hop neighbors exhibit a more generic and descriptive orientation, delving into a wide array of user experiences. The two clouds highlight the depth and diversity of discussions within the dataset. The study generated multiple word clouds by adjusting the maximum number of words displayed, consistently yielding the same interpretation.

Rather than naively collecting posts and comments from the initial 9 seed subreddits, the proposed framework explores all the subreddits where seed subreddits’ members have created a post as well. Hence, the method ends up collecting data from 27 additional subreddits. Before including the one-hop neighbors, the dataset contained 61,922 posts and 180,289 comments. After adding the one-hope neighbors, the number of posts increased to 4,271,890 and the number of comments was 8,719,303. The number of posts experienced a significant increase by a factor of 68.98, while the number of comments saw a notable increase by a factor of 48.36. This keeps in line with the observation that a common trait of subreddits is that most posts contain few or no comments.

Table 2 compares the topological features of the graph that contains only the 9 seed subreddits’ data vs the proposed graph. The graph exhibits a remarkable surge in both the quantity of nodes and edges. The average degree of the proposed graph is higher than the seed subreddits graph, but closer examination reveals that user nodes’ degrees remain the same. Instead, the average degree for subreddits shows a huge rise. This is attributed to the neighbors’ collection method. This part of the framework selects neighbors

where at least 1% of the seed subreddits' users were interacting and hence, favors popular subreddits. Additionally, results indicate that the density of the two graphs is comparable but the clustering coefficient of the proposed graph is less than that of the seed subreddit graph. This is expected behavior since numerous user-subreddit links have been added in the proposed graph and subreddit nodes are not directly connected by design.

Property	Seed subreddit graph	Proposed graph
Number of nodes	35,964	1,551,648
Number of edges	98,987	4,542,276
Density	$1.53 * 10^4$	$3.77 * 10^6$
Number of components	1	1
Average degree	5.50	5.85
Average degree (users)	4.83 (±14.85)	5.09 (±113.02)
Average degree (subreddits)	2,696.67 (±2,465.38)	32,888.41 (±83,086.90)
Clustering coefficient	0.15	0.11

Table 2. Network properties comparison between seed subreddits graph and proposed graph.

In summary, the cumulative findings suggest that inclusion of one hop neighbors greatly expands the amount of information being added to the graph while basic properties like user node degrees, degree distribution and clustering coefficient is maintained. This proves that adding neighborhood subreddits add value to the analysis and the proposed graph provides an in-depth perspective about how users interact with PPC technology.

6.2 To what extent is the information retained within the proposed graph? (RQ 2)

If the suggested graphing method effectively preserves user interaction information by storing interaction-based edges and textual information node attributes, the analysis of the graph can be used to successfully predict missing links within the graph. In Table 3, the outcomes of the link prediction task are presented. This comparative analysis involves evaluating the proposed graph against several alternatives, each of which has its unique characteristics.

Graph type	User-user edges		User-subreddit edges	
	Accuracy	F1-score	Accuracy	F1-score
Proposed heterogeneous graph	**91.1%**	**0.90**	**94.5%**	**0.95**
Graph with no textual embeddings	72.2%	0.75	89.0%	0.89
User-user graph with comment links only	87.1%	0.86	-	-
User-subreddit graph without comment links	-	-	90.0%	0.90
BERT model finetuned on text only	83.1%	0.76	92.3%	0.91

Table 3. Results for link prediction task. Graph performance is evaluated in predicting user-user and subreddit-user links.

First, the proposed graph is compared with one that contains no textual node attributes. Then it is juxtaposed with a user-subreddit graph that lacks comment information and a homogeneous user-user graph that exclusively consists of comment links. The user-user graph in the context of digital marketing is a widely employed method, wherein users are connected if they post in the same subreddit. However, it is crucial to acknowledge that in the dataset, several subreddits have a substantial number of users posting within them, such as 'r/business,' which boasts over 200,000 users contributing to the community. As a result, this configuration generates numerous complete subgraphs within the overall graph. To illustrate the magnitude of the challenge, creating an adjacency matrix to represent such a user-user graph with over 1.5 million nodes would necessitate a staggering 3.4 terabytes of computational storage space. Due to

hardware limitations, the study was unable to undertake a direct comparison with such a user-user graph. In contrast, the proposed graph is stored as an edge list, requiring a significantly more manageable 72.7 megabytes of storage space. Finally, the proposed approach is contrasted with a BERT model trained on user-user and user-subreddit texts without incorporating any graph structure data. Due to hardware limitations, the training phase for BERT utilized 100,000 pairs, with an additional 100,000 samples reserved for testing purposes.

The results of the analysis demonstrate that the proposed heterogeneous network, which incorporates users and subreddits as nodes, combined with BERT-based node textual embeddings, outperforms the three graph alternatives considered, achieving an accuracy rate of 91.1% and an F1-score of 0.90 for user-user link prediction, and 94.5% accuracy with a 0.95 F1-score for user-subreddit link prediction. The drop in performance for user-user link prediction is a reasonable outcome, considering that predicting user-user interactions is a more complex task. These interactions often involve shorter, less formal texts with limited context. Furthermore, the proposed graph exhibits substantial improvement over the baseline approaches. Specifically, an impressive 18.9% enhancement in user-user link prediction and a 5.5% boost in user-subreddit link prediction performance is observed and is primarily attributed to the inclusion of textual embeddings. The incorporation of a heterogeneous node structure also amplifies the model's capability to predict different types of links.

When comparing the proposed model to BERT, which lacks access to graph structure, a significant performance improvement is observed. Specifically, the model outperforms BERT by 8% in user-user link prediction. In the context of user-subreddit link prediction, the proposed model demonstrates a performance advantage of 2.2% over the LLM approach. There are two potential explanations for these findings. One possibility is that BERT's performance is hindered by the absence of graph structure. Alternatively, the disparity could be attributed to the training size for BERT. Due to the computational expense of training and fine-tuning the model, it is challenging to train the model on all links adequately, which may impact performance evaluation.

To conclude, the obtained results underscore the effectiveness of creating a heterogeneous graph and storing textual information as node attributes within the graph. They substantiate the robustness of the designed graphing methodology in terms of information preservation and the prediction of missing links, reinforcing the merits of the proposed approach in advancing network analysis and link prediction tasks.

6.3 How much value does contextual information contribute to the performance of the graph in the link prediction task? (RQ 3)

Table 4 provides an insightful analysis of the performance of graph neural network models in the context of the overall link prediction task when paired with various text vector generation models. It is important to note that this comparison involves the use of pre-trained versions of BERT, RoBERTa, and BERTweet, while both GloVe and Doc2Vec were trained on the data and employed as baselines.

The results indicate that BERT outshines all the baseline approaches. BERTweet, on the other hand, encounters a performance setback due to its shorter token length of 128. This limitation results in a significant loss of information when representing text as BERTweet embeddings. The superior performance of BERT can be attributed to its capacity to generate contextually aware embeddings, which enables it to outperform both GloVe and Doc2Vec. Furthermore, BERT exhibits a marginally better performance compared to RoBERTa, with only a 0.5% difference. It is worth noting that the evidence does not conclusively establish BERT's superiority over RoBERTa for this task. Hence, researchers are encouraged to consider both methods and determine which model aligns better with their specific objectives.

In summary, all the textual embedding methods consistently outperform a graph with no textual embeddings, showcasing a minimum improvement of 9.2%. These results unequivocally underscore the significant value that textual embeddings bring to the proposed graph, emphasizing the pivotal role of text-based information in enhancing the graph's predictive capabilities. It is noted that Doc2Vec performs

better than the FastText model. A possible explanation is that Doc2Vec is designed to capture document level information i.e., while considering the meaning of the words, it also takes broader sentence-based context into account. FastText focuses on just word embeddings and hence, is unable to generate effective document level vectors.

Text embedding model	Accuracy	F1-score
BERT	**92.1%**	**0.91**
BERtweet	90.1%%	0.89
RoBERTa	91.6%	**0.91**
GloVe	87.6%	0.82
Doc2Vec	88.0%	0.87
No textual embeddings	78.4%	0.81

Table 4. Results for link prediction task for graphs with different text embedding models. The evaluation of graph performance assesses its ability to predict missing links.

7 News Analytics Application

The applicability and robustness of the proposed methodology across at a different domain is achieved using an in-house dataset comprising news articles and corresponding comments spanning from June 2021 to June 2022, sourced from the archives of the New York Times. This dataset encompasses a substantial corpus, encompassing 6580 articles and 211,415 comments, covering a spectrum of topics including but not limited to entities such as Philadelphia, the European Union, and the Oscars, among others. Our focus is analyzing news articles about the Russia-Ukraine conflict. This experiment is initiated by selecting two seed articles, namely 'The Cancellation of Mother Russia Is Underway' and 'In Ukraine, It's Putin's Plan B vs. Biden's and Zelensky's Plan A'. Subsequently, a comprehensive list of users who commented on these articles was compiled and their immediate one-hop neighbor articles within the network were identified. Finally, there were 236 potentially relevant news articles. To streamline the analysis, a 0.75% user threshold was applied, reducing the number of one-hop neighbors to 55 articles. Higher and lower thresholds (1.5%, 1%, and 0.5%) were also experimented with, but it was found that thresholds above 0.75% led to significant information loss, while a threshold of 0.5% introduced excessive noise. At this stage, user intervention was introduced to manually select the top 28 one-hop neighbors most relevant to our analysis. Some of the one-hop news articles are 'Ukraine Is the First Real World War', 'In Ukraine, Putin's Gamble Has Failed', and 'How the U.S. Raced to Arm Ukraine Against Russia'. The articles closely aligned with the thematic focus on the Russia-Ukraine conflict.

Figure 5 illustrates the manually interpreted BERTopic-generated topics for both the seed articles and their corresponding one-hop neighbors, analyzed independently. In this visualization, all the comments from the articles were leveraged for topic modeling, excluding the text from the news articles themselves. A notable observation is the evident overlap between the topics emerging from the seed articles and their neighborhoods. For instance, discussions revolving around themes such as Putin, China, nuclear weapons, and sanctions against Russia are discernible in both groups. Moreover, the inclusion of one-hop neighbors unveils novel topics such as the China-Taiwan problem, the US military's involvement in the conflict, and the varying regional reactions or involvements in the conflict (e.g., Germany, Crimea, Chechnya, etc.). Additionally, users express concerns regarding various aspects, including the bombings, Russia's GDP, and the potential for chemical warfare. This analysis highlights how the proposed data collection method not only enriches the analysis of a particular topic but also demonstrates its versatility across various

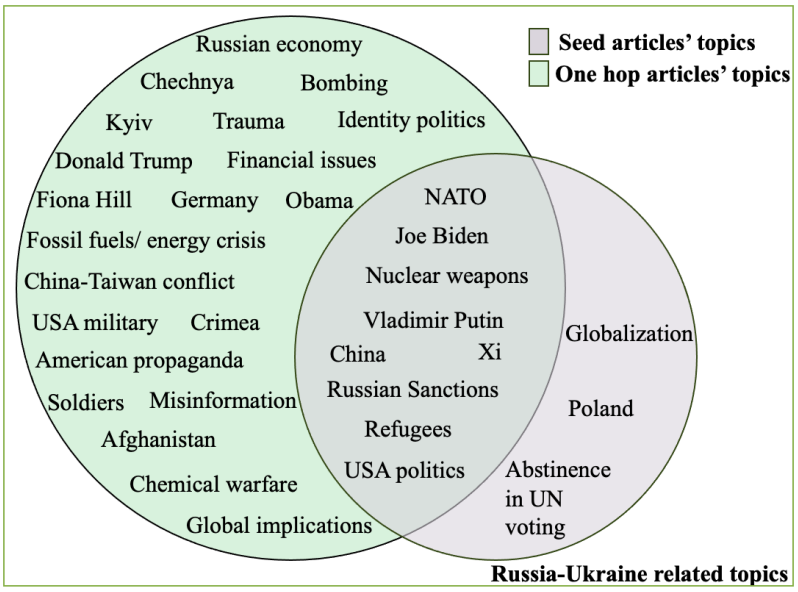

Figure 5. Manually interpreted topics generated for seed news articles and one-hop neighbors.

platforms beyond Reddit.

8 Conclusion

This study proposes an investigation into topic-related user experiences by analyzing interactions on social platforms like Reddit. Given a list of seed subreddits as input, the dataset is created by collecting all the posts and comments in these subreddits as well as from the one-hop neighbors that are topically related to the seed subreddits. the next step of the framework generates a static heterogeneous graph, containing subreddits and users as nodes and user interactions as edges. It incorporates BERT-based node attributes for textual information storage. The entire framework successfully captures users' interaction with minimal human intervention therefore it is cost-effective while providing a broad view of users' experience with a topic. The study employs a Pay-per-Click technology use case to demonstrate the effectiveness of the proposed method. The experiments show that one-hop neighborhood analysis significantly enhances the graph's topical value, revealing significant topic overlap among seed subreddit and one-hop neighbors. Neighborhood analysis also introduces new topics that were previously missing in seed subreddit discussions. Word cloud analysis highlights users sharing experiences with PPC in these communities. Link prediction results demonstrate the graph's effective data storage capability by achieving 92.1% overall accuracy and a 0.91 F1-score. The generalizability of the proposed method is demonstrated by using New York Times news articles and the Russia-Ukraine war as testing examples. It is important to note that the selected one-hop neighbors may not represent all aspects of the topic. In addition to this, the framework generates a static graph, but temporal information is not considered. Similarly, the topic evolution and precursors are not discussed. Future works can delve into overcoming these limitations.

Acknowledgments

The research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

We also thank Dr. Jumanah Alshehri for database support.

References

Agarwal, M. (2021). "A study on Pay-Per-Click advertising." *Asian Journal of Multidimensional Research* 10 (11), 618–624.

Alalwan, A. A., N. P. Rana, Y. K. Dwivedi, and R. Algharabat (2017). "Social media in marketing: A review and analysis of the existing literature." *Telematics and informatics* 34 (7), 1177–1190.

Aleksic, N., E. Pajic, P. Obradovic, W. Power, M. Mišic, and Z. Obradovic (2022). "Modelling subreddit interactions by activity overlap."

Alshehri, J., M. Stanojevic, E. Dragut, and Z. Obradovic (2021). "Stay on topic, please: aligning user comments to the content of a news article." In: *European Conference on Information Retrieval*. Springer, pp. 3–17.

Alzahrani, R. A. and M. Aljabri (2022). "AI-based techniques for Ad click fraud detection and prevention: Review and research directions." *Journal of Sensor and Actuator Networks* 12 (1), 4.

Amaya, A., R. Bach, F. Keusch, and F. Kreuter (2021). "New data sources in social science research: Things to know before working with Reddit data." *Social science computer review* 39 (5), 943–960.

Baumgartner, J., S. Zannettou, B. Keegan, M. Squire, and J. Blackburn (2020). "The pushshift reddit dataset." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 830–839.

Benslimane, S., J. Azé, S. Bringay, M. Servajean, and C. Mollevi (2023). "A text and GNN based controversy detection method on social media." *World Wide Web* 26 (2), 799–825.

Boettcher, N. (2021). "Studies of depression and anxiety using reddit as a data source: scoping review." *JMIR mental health* 8 (11), e29487.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017). "Enriching word vectors with subword information." *Transactions of the association for computational linguistics* 5, 135–146.

Bonifazi, G., E. Corradini, D. Ursino, and L. Virgili (2023). "Modeling, Evaluating, and Applying the eWoM Power of Reddit Posts." *Big Data and Cognitive Computing* 7 (1), 47.

Brambilla, M. and M. Gasparini (2019). "Brand community analysis on social networks using graph representation learning." In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pp. 2060–2069.

Desai, V. and B. Vidyapeeth (2019). "Digital marketing: A review." *International Journal of Trend in Scientific Research and Development* 5 (5), 196–200.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.

Flaherty, T., C. Domegan, and M. Anand (2021). "The use of digital technologies in social marketing: a systematic review." *Journal of Social Marketing* 11 (4), 378–405.

Ge, R., H. Zhao, and S. Zhang (2022). "Online Brand Community User Segments: A Text Mining Approach." *Frontiers in artificial intelligence* 5, 900775.

Gkikas, D. C. and P. K. Theodoridis (2019). "Artificial intelligence (AI) impact on digital marketing research." In: *Strategic Innovative Marketing and Tourism: 7th ICSIMAT, Athenian Riviera, Greece, 2018*. Springer, pp. 1251–1259.

Gori, M., G. Monfardini, and F. Scarselli (2005). "A new model for learning in graph domains." In: *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.* Vol. 2. IEEE, pp. 729–734.

Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794*.

Hessel, J., C. Tan, and L. Lee (2016). "Science, askscience, and badscience: On the coexistence of highly related communities." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 10. 1, pp. 171–180.

Hudson, S., L. Huang, M. S. Roth, and T. J. Madden (2016). "The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors." *International Journal of Research in Marketing* 33 (1), 27–41.

Hurtado, S., P. Ray, and R. Marculescu (2019). "Bot detection in reddit political discussion." In: *Proceedings of the fourth international workshop on social sensing*, pp. 30–35.

Janchevski, A. and S. Gievska (2019). "A study of different models for subreddit recommendation based on user-community interaction." In: *ICT Innovations 2019. Big Data Processing and Mining: 11th International Conference, ICT Innovations 2019, Ohrid, North Macedonia, October 17–19, 2019, Proceedings 11*. Springer, pp. 96–108.

Jiang, Z. P., S. I. Levitan, J. Zomick, and J. Hirschberg (2020). "Detection of mental health from reddit via deep contextualized representations." In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pp. 147–156.

Kapoor, K. K., Y. K. Dwivedi, and N. C. Piercy (2016). "Pay-per-click advertising: A literature review." *The Marketing Review* 16 (2), 183–202.

Krishen, A. S., Y. K. Dwivedi, N. Bindu, and K. S. Kumar (2021). "A broad overview of interactive digital marketing: A bibliometric network analysis." *Journal of Business Research* 131, 183–195.

Law, E. L.-C. and P. Van Schaik (2010). *Modelling user experience–An agenda for research and practice*.

Le, Q. and T. Mikolov (2014). "Distributed representations of sentences and documents." In: *International conference on machine learning*. PMLR, pp. 1188–1196.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692*.

Medvedev, A. N., R. Lambiotte, and J.-C. Delvenne (2019). "The anatomy of Reddit: An overview of academic research." *Dynamics on and of Complex Networks III: Machine Learning and Statistical Physics Approaches 10*, 183–204.

Nguyen, D. Q., T. Vu, and A. T. Nguyen (2020). "BERTweet: A pre-trained language model for English Tweets." *arXiv preprint arXiv:2005.10200*.

Oliveira, N. and J. P. Marques dos Santos (2022). "Online Brand Community Characterization with Engagement and Social Network Analysis (SNA) for Marketing Communication: The Subreddit r/intel." In: *Marketing and Smart Technologies: Proceedings of ICMarkTech 2021, Volume 2*. Springer, pp. 577–591.

Prabowo, N. A., B. Pujiarto, F. S. Wijaya, L. Gita, and D. Alfandy (2021). "Social network analysis for user interaction analysis on social media regarding e-commerce business." *International Journal of Informatics and Information Systems* 4 (2), 95–102.

Proferes, N., N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer (2021). "Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics." *Social Media+ Society* 7 (2), 20563051211019004.

Shankar, V. and S. Parsana (2022). "An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing." *Journal of the Academy of Marketing Science* 50 (6), 1324–1350.

Yang, L., Q. Cheng, and S. Tong (2015). "Empirical study of eWOM's influence on consumers' purchase decisions." *The Strategies of China's Firms*, 123–135.

Zhu, L. and J. Lv (2023). "Review of studies on user research based on EEG and eye tracking." *Applied Sciences* 13 (11), 6502.

Zhu, Y., Y. Wang, J. Wei, and A. Hao (2023). "Effects of vividness, information and aesthetic design on the appeal of pay-per-click ads." *Journal of Research in Interactive Marketing* 17 (5).